Convergence of Dynamic Programming in Reinforcement Learning

Seok Won Lee

ijleesw@gmail.com

1 Convergence of Value Iteration

Value function $V : S \to \mathbb{R}$에 대해, Bellman Operator $\mathcal{B} : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$를

$$\mathcal{B}V(s) = R(s) + \gamma \max_a \sum_{s'} P_{sa}(s')V(s')$$

로 정의하자.

즉, $\mathcal{B}V$는 value iteration에 의해 update된 value function을 의미한다.

Lemma 1.1. V^*는 $\mathcal{B}V$로 update되어도 자기 자신이다.

Proof. Bellman Equation에 의해

$$V^*(s) := \max_\pi V^\pi(s) = R(s) + \max_a \left(\gamma \sum_{s'} P_{sa}(s')V^*(s') \right)$$

이므로, $\mathcal{B}V^*(s) = V^*(s)$ 가 성립한다.

□

Lemma 1.2. 두 value function이 value iteration에 의해 update되면, $V_1(s)$와 $V_2(s)$의 차이의 최댓값이 일정 비율$(= \gamma)$ 이상 줄어든다.

Proof. 임의의 두 value function V_1, V_2에 대해,

$$\begin{aligned}
|\mathcal{B}V_1(s) - \mathcal{B}V_2(s)| &= \gamma |\max_a \sum_{s'} P_{sa}(s')V_1(s') - \max_a \sum_{s'} P_{sa}(s')V_2(s')| \\
&\le \gamma \max_a |\sum_{s'} P_{sa}(s')V_1(s') - \sum_{s'} P_{sa}(s')V_2(s')| \\
&= \gamma \max_a \sum_{s'} P_{sa}(s')|V_1(s') - V_2(s')| \\
&\le \gamma \max_a |V_1(s) - V_2(s)| \\
&= \gamma ||V_1 - V_2||_\infty
\end{aligned}$$

이므로, $||\mathcal{B}V_1 - \mathcal{B}V_2||_\infty \leq \gamma||V_1 - V_2||_\infty$ 가 성립한다.

□

Theorem. V^*와 임의의 V 간의 value의 차이의 최댓값은 iteration을 반복하면 0에 수렴한다.

Proof. Lemma 1.1.과 1.2.에 의해

$$\begin{aligned} ||V_k - V^*||_\infty &= ||\mathcal{B}V_{k-1} - \mathcal{B}V^*||_\infty \leq \gamma||V_{k-1} - V^*||_\infty \\ &= \gamma||\mathcal{B}V_{k-2} - \mathcal{B}V^*||_\infty \leq \gamma^2||V_{k-2} - V^*||_\infty \\ &\leq \cdots \\ &\leq \gamma^k||V_0 - V^*||_\infty \quad \text{as } k \to \infty \end{aligned}$$

이므로, value iteration은 수렴한다.

□

2 Convergence of Policy Iteration

Lemma 2.1. (Banach Fixed-Point Theorem)
$\mathbb{R}^n$에서 정의된 contraction map (= 점들 사이의 거리를 줄이는 함수)은 unique fixed-point를 갖는다.

Proof. $\mathcal{B} : \mathbb{R}^n \to \mathbb{R}^n$가 고정된 $\gamma \in [0,1)$에 대해 $||\mathcal{B}(x) - \mathcal{B}(y)|| \leq \gamma||x - y|| \quad \forall\, x, y \in \mathbb{R}^n$이라 하자. 임의의 $x \in \mathbb{R}^n$에 대해 $x^* = \lim_{n\to\infty} \mathcal{B}^n(x)$라고 하면 x^*는 fixed-point이다. 따라서 fixed-point는 존재한다.

다른 한편 삼각부등식에 의해 다음이 성립한다:

$$\begin{aligned} ||\mathcal{B}(x) - \mathcal{B}(y)|| &\leq ||x - \mathcal{B}(x)|| + ||\mathcal{B}(x) - \mathcal{B}(y)|| + ||\mathcal{B}(y) - y|| \\ &\leq ||x - \mathcal{B}(x)|| + \gamma||x - y|| + ||\mathcal{B}(y) - y||. \end{aligned}$$

위 부등식을 정리하면

$$||x - y|| \leq \frac{||\mathcal{B}(x) - x|| + ||\mathcal{B}(y) - y||}{1 - \gamma}$$

를 얻는다. 따라서 $\mathcal{B}(x) = x$, $\mathcal{B}(y) = y$라면 $||x - y|| = 0$이므로 $x = y$이고, fixed-point는 유일하다.

□

Policy function $\pi_k : S \to A$에 대해, $\pi_{k+1} : S \to A$를

$$\pi_{k+1} := \operatorname*{argmax}_a \sum_{s'} P_{sa}(s')V^{\pi_k}(s')$$

로 정의하자.

그리고 π_k에 의해 주어진 value function V^{π_k}에 대해, Bellman Operator $\mathcal{B}^{\pi_i} : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$를

$$\mathcal{B}^{\pi_i} V^{\pi_k}(s) = R(s) + \gamma \sum_{s'} P_{s\pi_i(s)}(s') V^{\pi_k}(s')$$

로 정의하자.

그러면 모든 s에 대해 다음이 성립한다:

$$\begin{aligned}
\mathcal{B}^{\pi_{k+1}} V^{\pi_k}(s) &= R(s) + \gamma \sum_{s'} P_{s\pi_{k+1}(s)}(s') V^{\pi_k}(s') \\
&= R(s) + \gamma \max_a \sum_{s'} P_{sa}(s') V^{\pi_k}(s') \\
&= \mathcal{B} V^{\pi_k}(s) \\
&\geq R(s) + \gamma \sum_{s'} P_{s\pi_k(s)}(s') V^{\pi_k}(s') \\
&= V^{\pi_k}(s).
\end{aligned}$$

Lemma 2.2 위에서 정의한 Bellman Operator $\mathcal{B}^{\pi_i}$는 contraction map이다.

Proof. Lemma 1.2.의 증명과 마찬가지로,

$$\begin{aligned}
||\mathcal{B}^{\pi_i} V_1 - \mathcal{B}^{\pi_i} V_2||_\infty &= \max_s |\mathcal{B}^{\pi_i} V_1(s) - \mathcal{B}^{\pi_i} V_2(s)| \\
&= \gamma \max_s |\sum_{s'} P_{s\pi_i(s)}(s') V_1(s') - \sum_{s'} P_{s\pi_i(s)}(s') V_2(s')| \\
&\leq \gamma \max_a |\sum_{s'} P_{sa}(s') V_1(s') - \sum_{s'} P_{sa}(s') V_2(s')| \\
&= \gamma \max_a \sum_{s'} P_{sa}(s') |V_1(s') - V_2(s')| \\
&\leq \gamma \max_a |V_1(s) - V_2(s)| \\
&= \gamma ||V_1 - V_2||_\infty
\end{aligned}$$

이다.

□

Lemma 2.3. Policy iteration이 시행되면, update된 value function으로 구한 기댓값 $V^{\pi_{k+1}}(s)$가 이전 value function의 기댓값 $V^{\pi_k}(s)$보다 같거나 크다.

Proof. 우선 $\mathcal{B}^{\pi_{k+1}}V^{\pi_k}(s) \leq V^{\pi_k}(s)$ 에서

$$\begin{aligned}(\mathcal{B}^{\pi_{k+1}})^{m+1}V^{\pi_k}(s) &= R(s) + \gamma \sum_{s'} P_{s\pi_{k+1}(s)}(s')(\mathcal{B}^{\pi_{k+1}})^m V^{\pi_k}(s') \\ &\geq R(s) + \gamma \sum_{s'} P_{s\pi_{k+1}(s)}(s')(\mathcal{B}^{\pi_{k+1}})^{m-1} V^{\pi_k}(s') \\ &= (\mathcal{B}^{\pi_{k+1}})^m V^{\pi_k}(s) \\ &\textit{if} \quad (\mathcal{B}^{\pi_{k+1}})^m V^{\pi_k}(s') \geq (\mathcal{B}^{\pi_{k+1}})^{m-1} V^{\pi_k}(s') \quad \forall\ s' \in S,\end{aligned}$$

$$\begin{aligned}(\mathcal{B}^{\pi_{k+1}})^2 V^{\pi_k}(s) &= R(s) + \gamma \sum_{s'} P_{s\pi_{k+1}(s)}(s')\mathcal{B}^{\pi_{k+1}} V^{\pi_k}(s') \\ &\geq R(s) + \gamma \sum_{s'} P_{s\pi_{k+1}(s)}(s') V^{\pi_k}(s') \\ &= \mathcal{B}^{\pi_{k+1}} V^{\pi_k}(s)\end{aligned}$$

이므로, $(\mathcal{B}^{\pi_{k+1}})^{n+1}V^{\pi_k}(s) \geq (\mathcal{B}^{\pi_{k+1}})^n V^{\pi_k}(s) \quad \forall\ n \geq 1$ 임을 알 수 있다.

다른 한편, Lemma 2.2.에 의해 $\mathcal{B}^{\pi_{k+1}}$ contraction map이며, Lemma 2.3.에 의해 unique fixed point $V^{\pi_{k+1}}$를 갖는다.

따라서 $V^{\pi_k}(s) \leq \mathcal{B}^{\pi_{k+1}}V^{\pi_k}(s)$의 양변에 $\mathcal{B}^{\pi_{k+1}}$를 반복적으로 취해주면

$$V^{\pi_k}(s) \leq \mathcal{B}^{\pi_{k+1}}V^{\pi_k}(s) \leq (\mathcal{B}^{\pi_{k+1}})^2 V^{\pi_k}(s) \leq \cdots \leq \lim_{n\to\infty} (\mathcal{B}^{\pi_{k+1}})^n V^{\pi_k}(s) \leq V^{\pi_{k+1}}(s)$$

가 성립한다.

□

Theorem. Policy iteration을 사용하면 유한 번 내에 optimal policy π^*를 구할 수 있다.

Proof. Policy function은 유한 개 존재하는데, Lemma 2.3.에 의해 $k \geq 1$일 때 V^{π_k}는 단조증가한다. 따라서 $V^{\pi_N} = V^{\pi_{N+1}}$인 $N \in \mathbb{N}$이 존재한다. 그리고 V^{π_N}에 대해 다음 식이 성립한다:

$$V^{\pi_N} = V^{\pi_N+1} = \mathcal{B}^{\pi_{N+1}}V^{\pi_{N+1}} = \mathcal{B}^{\pi_{N+1}}V^{\pi_N} = \mathcal{B}V^{\pi_N}.$$

Lemma 1.2에 의해 $\mathcal{B}$는 contraction map이며, Lemma 2.1.에 의해 contraction map은 unique fixed-point를 갖는다. In particular, V^{π_N}이 $\mathcal{B}$의 unique fixed-point이다.

따라서 위에서 찾은 V^{π_N}이 optimal value function V^*이며, π_N이 optimal policy function π^*가 된다. 즉, policy iteration을 사용하면 유한 번 내에 optimal policy π^*를 구할 수 있다.

□